ESTIMATION AND POPULATION STATISTICS

Focus:

The objective of this exercise is to estimate a population of trees using random sampling, and it is preferred that the final estimate has a small standard deviation.

Overview:

Sometimes when it is necessary to know the size of a large population or collection of things, we are unable to actually count the members of the group. Either the number of individual items in the population is too large or we do not have physical access to the population. In such cases, estimation can be a useful tool. Different methods of estimation are available and can be applied depending on the type of population.

The most common and straightforward type of estimation is global random sampling. (see below) A slightly more complicated estimation method is stratified random sampling (see below), and this technique can produce more accurate estimations than the other by yielding smaller errors.

Procedure:

Find the experiment "**Count the trees**" in the VIRTUAL LABORATORY at **http://www.jhu.edu/~virtlab**.

In this exercise you are expected to bid for the timber rights on a large plot of isolated land. Since timber value is assumed to be proportional to the number of trees, a tree count can be converted into a bid price. So, to make a bid, you need to estimate how many trees are on the plot. If your estimate is too low, another more accurate bidder will bid higher and win the logging rights. If your estimate is too high (i.e., you offer too much money for the logging rights), your company could go bankrupt.

The area of interest is given as a Landsat image in false color—the redder the color, the more trees. Every pixel within the field contains its own complement of trees. Clicking on a pixel will produce an area showing trees. Clicking anywhere within this area will increment a counter and leave a red dot, so counting trees is fairly routine. But there are a lot of pixels. . .

One of the complications of the exercise is that the spatial distribution of the trees is not homogeneous, i.e., some areas have heavy concentrations of trees while others have light concentrations. Depending on how different these areas are in tree concentration will determine whether global random sampling or stratified sampling will yield the better estimate. It would be nice to get an estimate whose standard deviation is less than 15 percent.

The most instructive plan is to carry out the exercise using stratified sampling (See below.) You'll have to guess how many sample points to start with (maybe 20 points in each stratum or sub-area). Then, to convince yourself that stratified sampling can be a good idea, recalculate your data as if it had been taken with global random sampling. Compare the differences. If your standard deviations are too large you'll have to take more data. You're interested in a standard deviation of less than 15% of your estimated total.

Global random sampling:

This is very simple. Randomly select N pixels from the entire image; count the trees in each of those pixels; find the average number of trees per pixel; find the standard deviation of that estimate; calculate the standard deviation of the mean. If the standard deviation of the mean is too large, e.g., greater than 15% of the total number of trees, increase N. (The standard deviation is proportional to $1/\sqrt{N}$.) The total number of trees is the average value per pixel times the number of pixels *P*.

The average number of trees/pixel is $\overline{T} = \frac{1}{N}\sum_{i=1}^{N} T_i$, where T_i is the number of trees in the i^{th} pixel.

The standard deviation of $\overline{T}$ is $\sigma_{\overline{T}} = \frac{\sigma}{\sqrt{N}}$, where $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(T_i - \overline{T})^2}$.

The answer you want is $T_{tot} = P\overline{T}$ with standard deviation $\sigma_{T_{tot}} = P\sigma_{\overline{T}}$.

Stratified random sampling:

This has more steps but could be a much better estimation procedure. Here, divide the image into *M* homogeneous sub-areas, i.e., sub-areas which are similar in tree density. Three or four areas should be adequate. Next, determine (or estimate) how many pixels there are in each of the *M* areas. Call these p_j for $j = 1, M$. The total number of pixels $P = \sum_{j=1}^{M} p_j$ should equal the total number of pixels in the image.

The average number of trees/pixel in the jth sub-area is $\overline{T}_j = \frac{1}{N_j}\sum_{i=1}^{N_j} T_{i,j}$,where $T_{i,j}$ is the number of trees in the i^{th} pixel in the j^{th} sub-area. N_j is the total number of sampled pixels in the j^{th} sub-area.

The standard deviation of $\overline{T}_j$ is $\sigma_{\overline{T}_j} = \frac{\sigma_j}{\sqrt{N_j}}$, where $\sigma_j = \sqrt{\frac{1}{N_j}\sum_{i=1}^{N_j}(T_{i,j} - \overline{T}_j)^2}$.

The total number of trees in the image is $T_{tot} = \sum_{j=1}^{M} p_j \overline{T}_j$.

Let $W_j = \frac{p_j \overline{T}_j}{T_{tot}}$. Then, the standard deviation of T_{tot} is $\sigma_T = \sqrt{\sum_{j=1}^{M} W_j^2 \sigma_{\overline{T}_j}^2}$

Again, the answer you want is T_{tot} , the total number of trees, and its standard deviation $P\sigma_T$.

The reason this sampling technique might be better than the non-stratified one is because the standard deviations of each sub-area will be smaller than the standard deviation of the area as a whole.

Write-up:

In your report make sure you tabulate your data. Show your calculations. Make sure you report your estimate for the total number of trees in the area and the standard deviation of that estimate. Explain how you subdivided the area into sub-areas. Explain how you estimated the areas of each of the sub-areas. Also, explain how you "randomly" chose your sample points.